Bloque III: El nivel de transporte

Tema 8: Intercambio de datos TCP

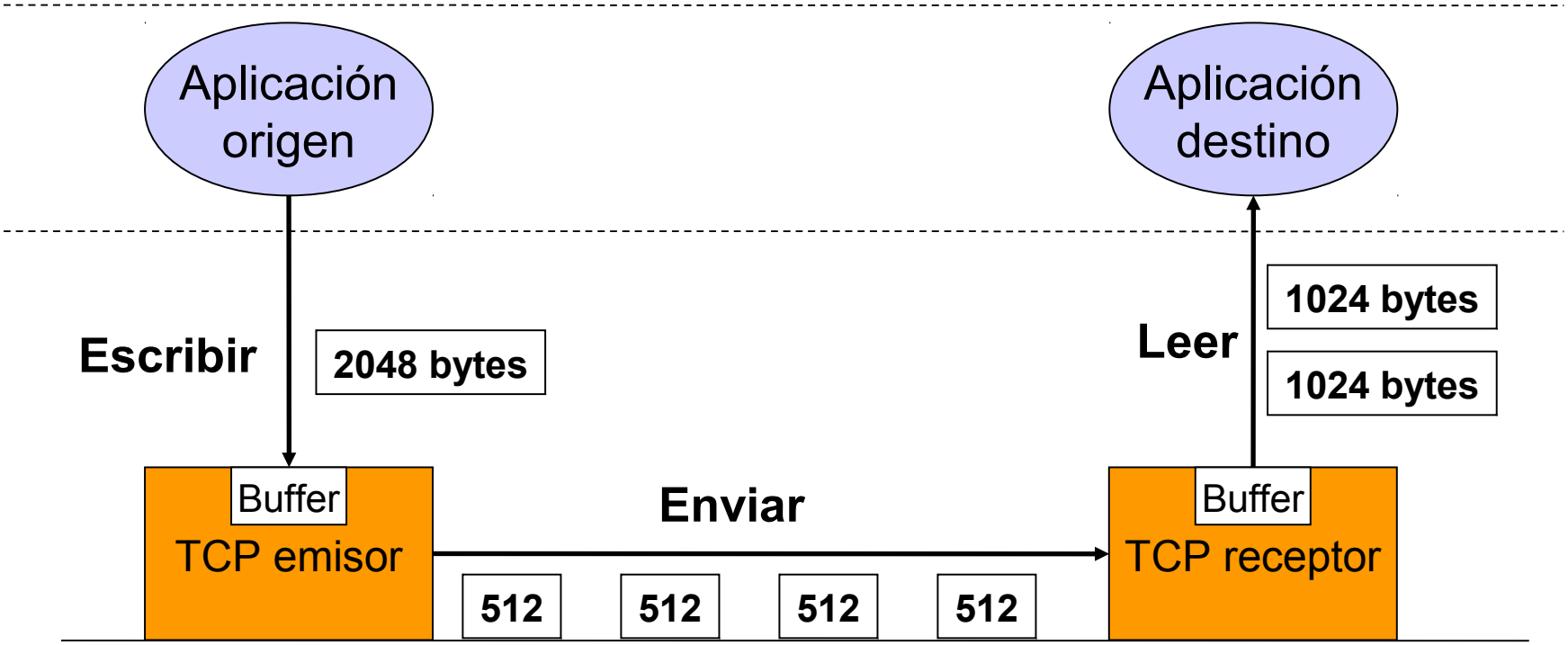
Índice

- Bloque III: El nivel de transporte
  - Tema 8: Intercambio de datos TCP
    - Introducción
    - Flujo de datos interactivo
      - ACKs retardados
      - Algoritmo de Nagle
    - Flujo de datos no interactivo
      - Control de flujo
      - Control de congestión
    - Temporizador de keepalive

- **Lecturas recomendadas:**
  - Capítulo 3, secciones 3.4, 3.5.3, 3.5.4, 3.5.5, 3.6.1 y 3.7 de “Redes de Computadores: Un enfoque descendente”. James F. Kurose, Keith W. Ross. Addison Wesley.
  - Capítulos 14, 15 y 17 de “TCP/IP Illustrated, Volume 1: The Protocols”, W. Richard Stevens, Addison Wesley.

TCP: Introducción

- En TCP se consideran dos tipos de tráfico de datos:
  - **Interactivo**: gran número de segmentos de pequeño tamaño (menos de 10 bytes). Por ejemplo: telnet, ssh.
  - **No Interactivo**: segmentos de gran tamaño, normalmente el máximo permitido por las limitaciones de la red. Por ejemplo: HTTP, FTP, e-mail.



TCP: Introducción

- Para implementar la fiabilidad, se basa en el modelo ARQ retroceder N:
  - Es un protocolo de ventana deslizante.
  - Los ACKs son acumulativos y positivos.
- Con algunos matices:
  - Cuando el receptor recibe un paquete fuera de orden:
    - No lo descarta, lo almacena en el buffer.
    - Envía un ACK del último paquete correcto.
  - Retransmisión rápida: si el emisor recibe tres ACKs repetidos → Retransmite (sólo) el paquete siguiente (al número de ACK).
  - El emisor mantiene un temporizador por cada grupo de paquetes enviado.

- En el RFC 2018 se propone un receptor TCP con confirmación selectiva (SACK).

TCP: Introducción

- Al tiempo de espera antes de retransmitir se le denomina: Retransmission Timeout (**RTO**).
- ¿Cuánto vale el RTO? Depende … del RTT (Round-Trip Time).
- ¿Cuánto vale el RTT? Depende … es variable a lo largo de la conexión.
- En una conexión TCP se estima el valor del RTT:
  - Cuando se envía un segmento, se mide el tiempo que tarda en recibirse su ACK.
  - Sólo se mide un RTT a la vez.
  - Si es un segmento retransmitido → No se calcula el RTT. ¿Por qué?
- Se calcula una media del RTT y su desviación típica (RTTDesv).
- El RTO = RTTEstimado + 4 * RTTDesv.
- Se usa un reloj hardware común para medir el RTT → La granularidad depende de la implementación (en linux es 1 ms).

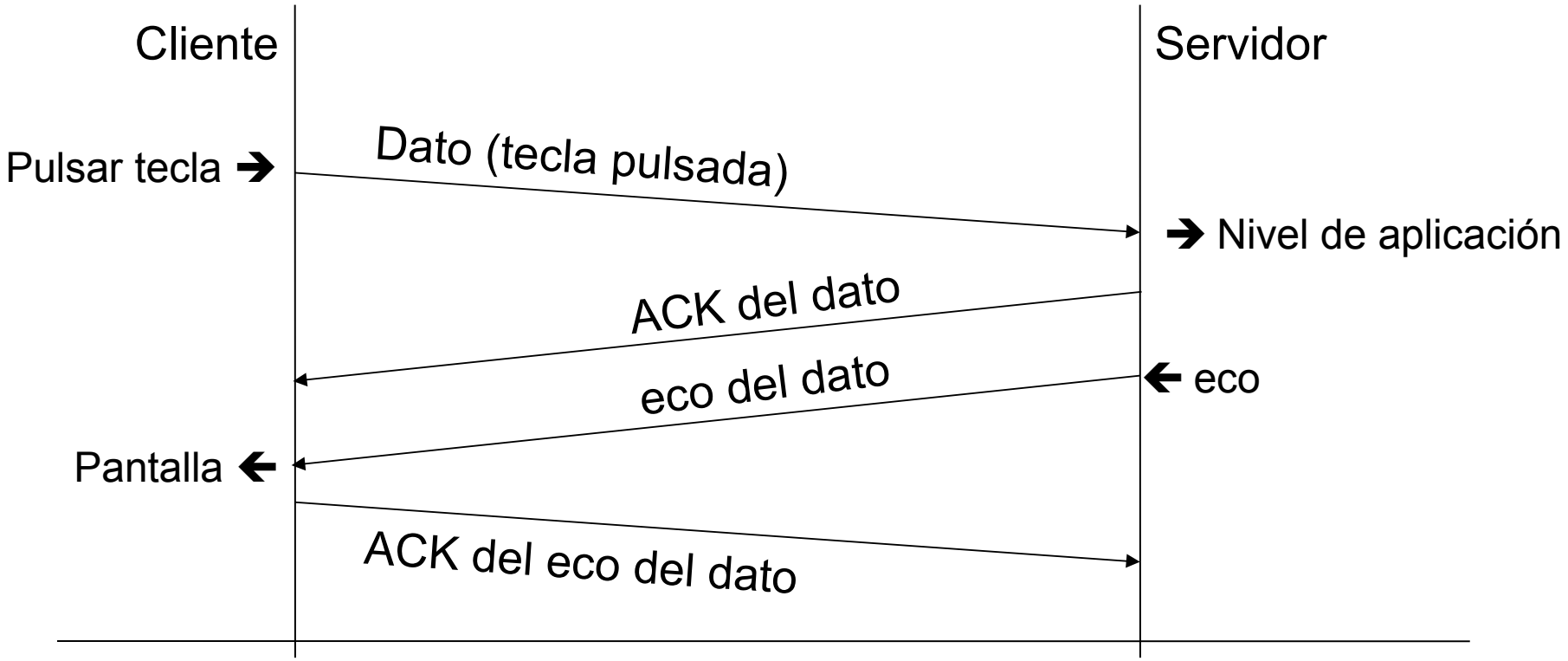
TCP: Introducción

- Opción de Timestamp (TSOPT): 10 bytes
  - Campo Timestamp Value (TSval): el emisor indica el valor de su reloj en el momento de la transmisión.
  - Campo Timestamp Echo Retry (TSecr): el receptor copia el TSV en el segmento de respuesta.
- Estimación del RTT con TSOPT:
  - El emisor indica el TSval en los segmentos que envía.
  - Al preparar la respuesta (ACK), el receptor copia el TSval en el campo TSecr.
  - El emisor, al recibir el ACK, comprueba el reloj del sistema, le resta el TSecr y tiene la estimación del RTT.
- Se controla en /proc/sys/net/ipv4/tcp_timestamps (por defecto activada).

- Para ver las estimaciones: *ip route show cache*

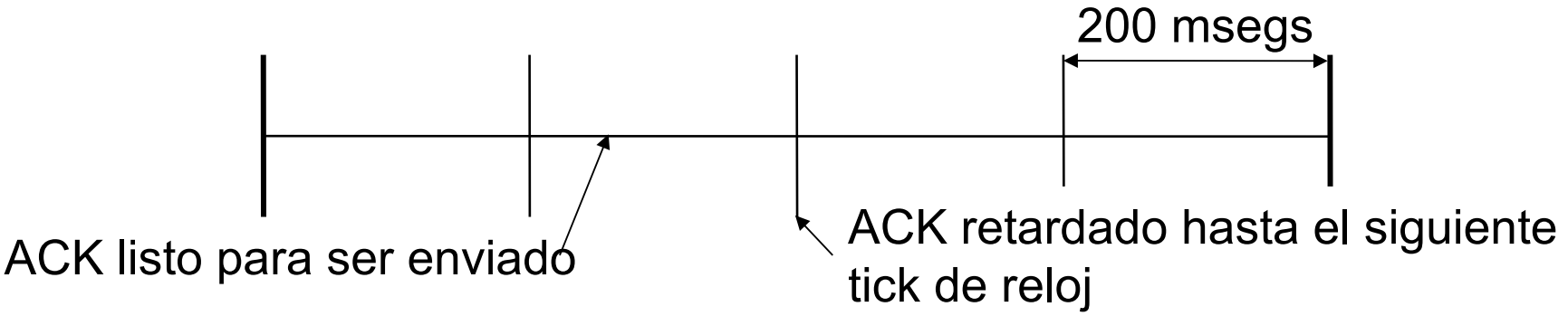
Flujo de datos interactivo

- Funcionamiento del ssh:
  - Envío de la tecla pulsada por el cliente
  - ACK de la tecla pulsada por el cliente
  - Eco de la tecla desde el servidor
  - ACK del eco

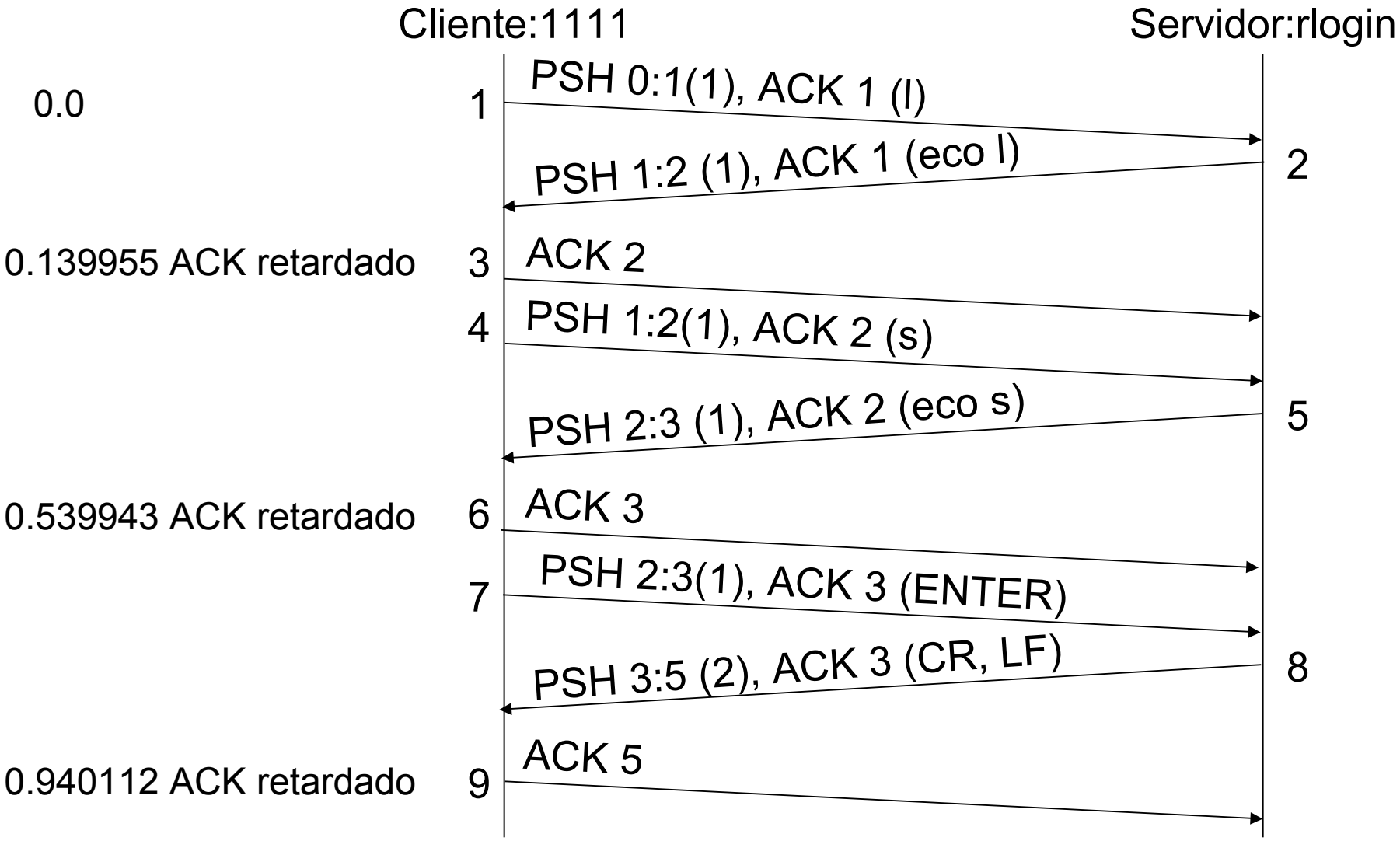


Flujo de datos interactivo

- ACKs retardados:
  - Objetivo: enviar el ACK + eco en un único datagrama.
- TCP no envía el ACK inmediatamente al recibir el dato, sino que retarda la salida del ACK esperando un tiempo para ver si hay datos para enviarlos con el propio ACK.
- El tiempo de espera es de 200 mseg:
  - No en valor absoluto, sino que se utiliza un reloj que da ticks cada 200 mseg.
- Tanto en el cliente como en el servidor se utilizan los acks retardados.



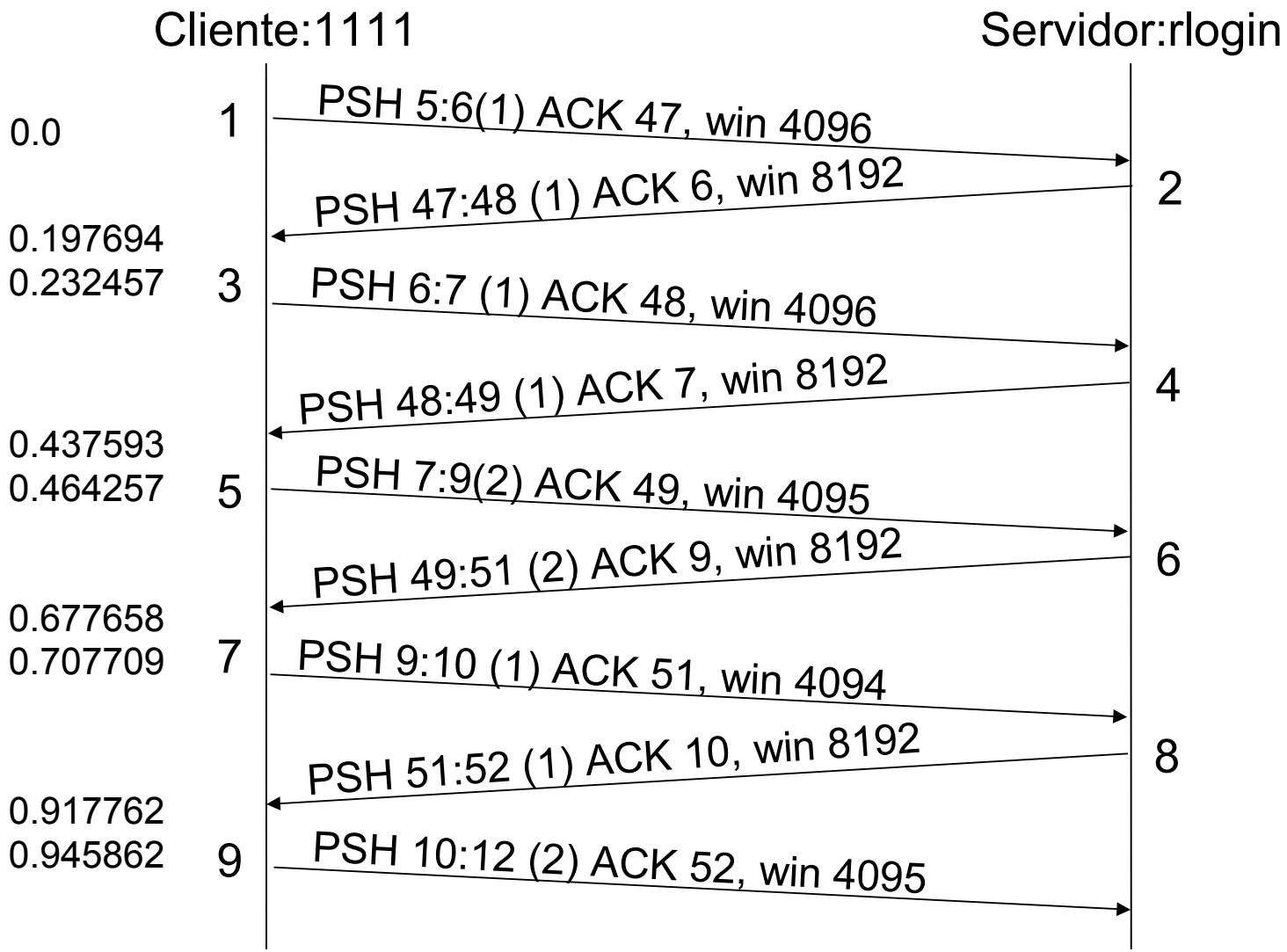
ACKs retardados



Flujo de datos interactivo

- El tráfico interactivo genera gran cantidad de paquetes de tamaño muy pequeño (1 byte datos + 20 cab. TCP + 20 cab. IP = 41 bytes) → **tinygrams**
  - En las redes de área local no presenta ningún problema
  - En las WAN supone una gran sobrecarga para la red.
- Algoritmo de Nagle:
  - Pretende resolver este problema, y se aplica en una conexión TCP de tráfico interactivo en una red de área extensa.
  - Enunciado: "*Una conexión TCP puede tener un único segmento pequeño que no haya sido confirmado. No se pueden enviar otros segmentos hasta recibir un ACK. En cambio, esos datos se almacenan y son enviados por TCP al llegar el ACK*."
  - Es auto-ajustable: cuanto más rápido lleguen los ACKs → más rápido se enviarán los datos.

Algoritmo de Nagle



Algoritmo de Nagle

- Se usan 5 segmentos para enviar 7 bytes de datos.
- ACK = TCP (nivel de transporte)
- Win = nivel de aplicación

- El algoritmo de Nagle convierte a TCP en un protocolo de parada y espera.
- En algunos casos puede no interesar:
  - Por ejemplo, al utilizar las XWindows es necesario enviar mensajes pequeños (movimientos del ratón) sin retardos para conseguir una interacción real.
  - En este caso, el algoritmo de Nagle puede introducir retardos adicionales al ejecutar una aplicación interactiva a través de una WAN, al generar eventos (pulsaciones, movimientos ratón, …) que producen datos multi-byte.
- Se puede controlar:
  - setTcpNoDelay(boolean) de la clase Socket.

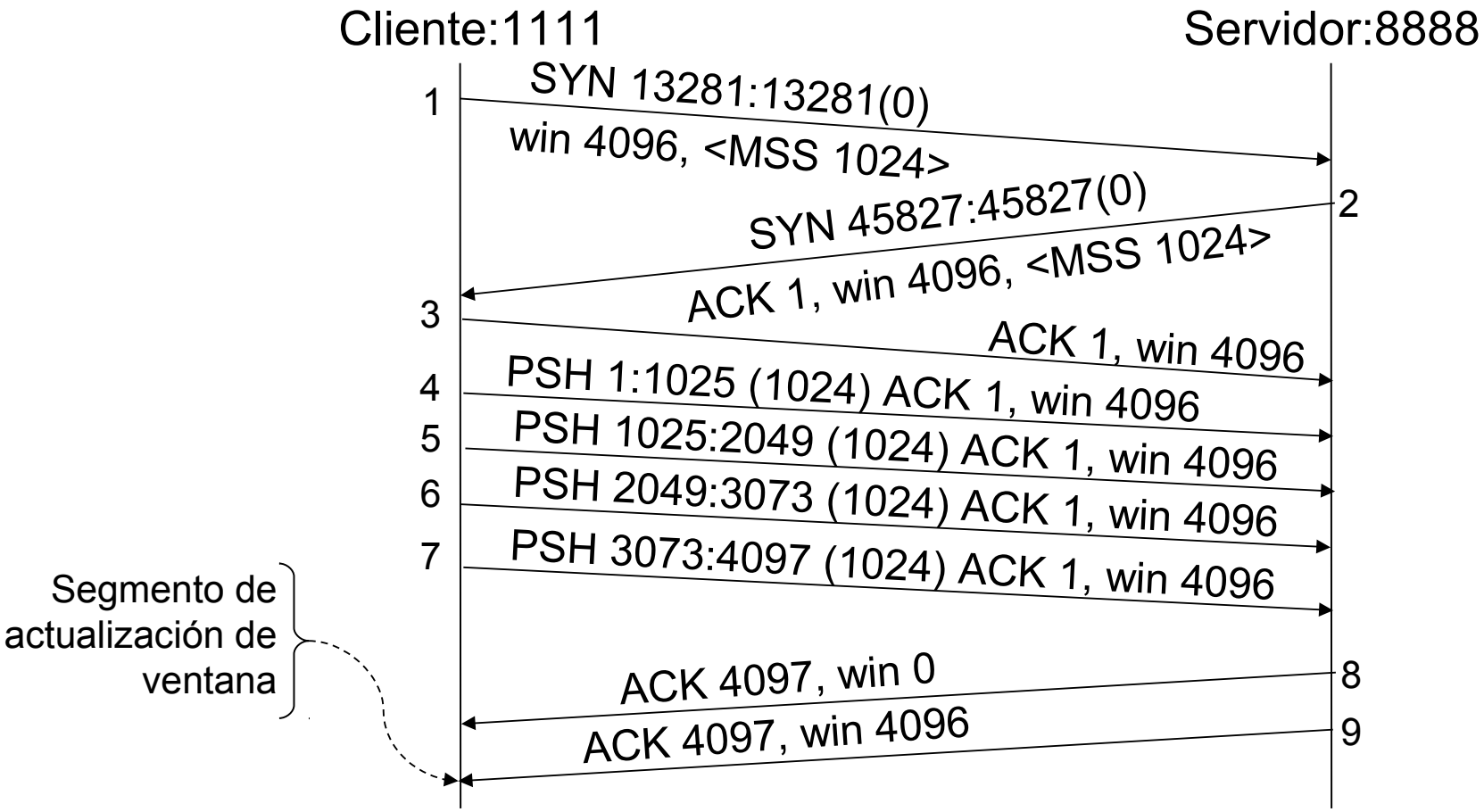
Flujo de datos no interactivo

- También llamado flujo de datos en masa (*bulk data flow*).
- En este tipo de tráfico se generan “pocos” segmentos, pero de gran tamaño.
- El principal problema a resolver en este tipo de tráfico es el **control de flujo**:
  - Evitar que un emisor rápido sature a un receptor lento.

- TCP utiliza una ventana deslizante: permite al emisor enviar múltiples paquetes antes de parar y esperar por el ACK, lo que da una mayor rapidez a este tipo de tráfico.
- Con un protocolo de ventana deslizante no es necesario confirmar todos los paquetes recibidos, sino que se pueden confirmar varios paquetes simultáneamente → ACKs acumulativos.

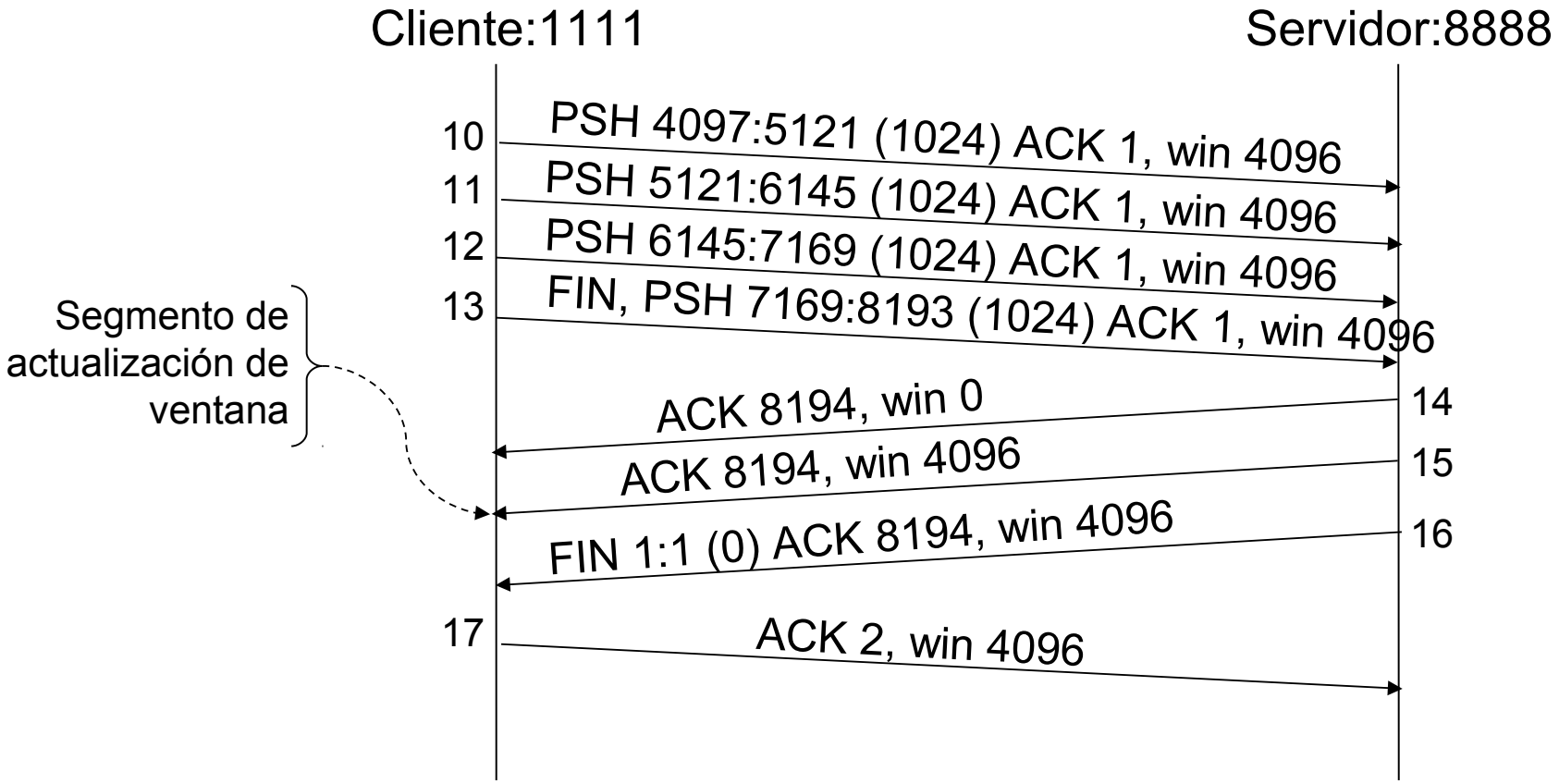
Control de flujo

- Emisor rápido, receptor lento



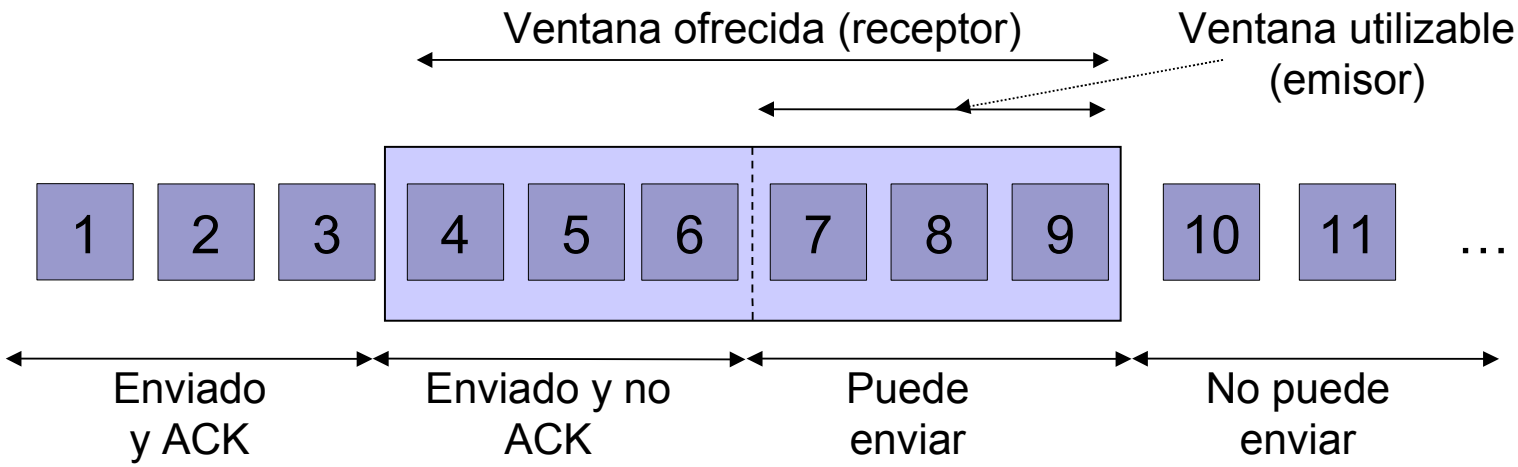
Control de flujo

- Emisor rápido, receptor lento (continuación)



Control de flujo

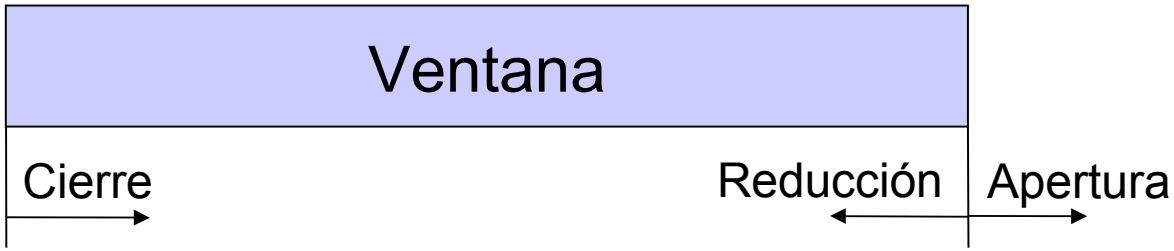
- Funcionamiento de la ventana deslizante:



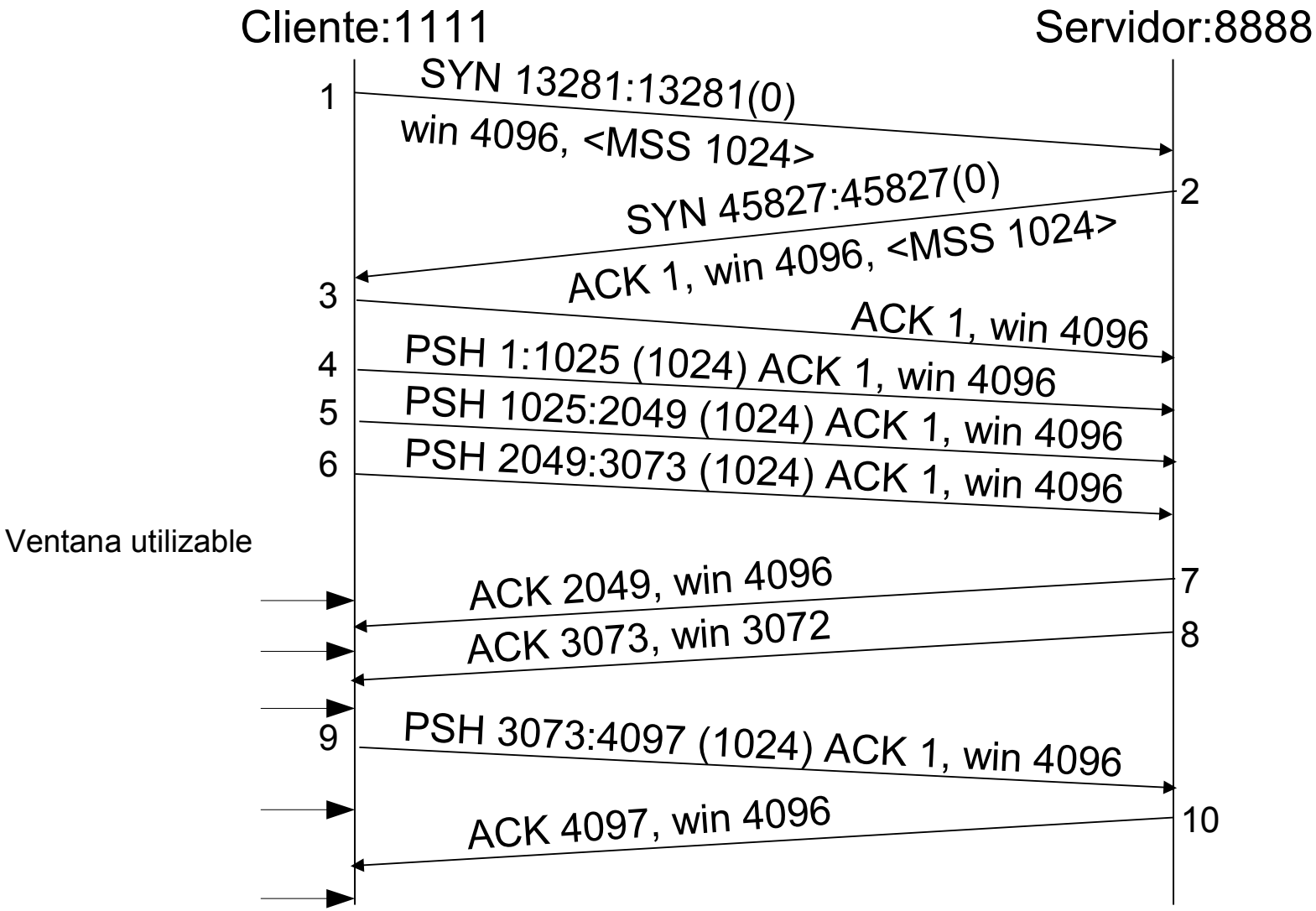
- **Ventana ofrecida** (receptor): número de bytes que indica el receptor que puede recibir en el momento de enviar el paquete (= win).
- **Ventana utilizable** (emisor): número de bytes que se pueden enviar inmediatamente.
  - Ventana utilizable = win – (sig. byte enviar – último ACK).

Control de flujo

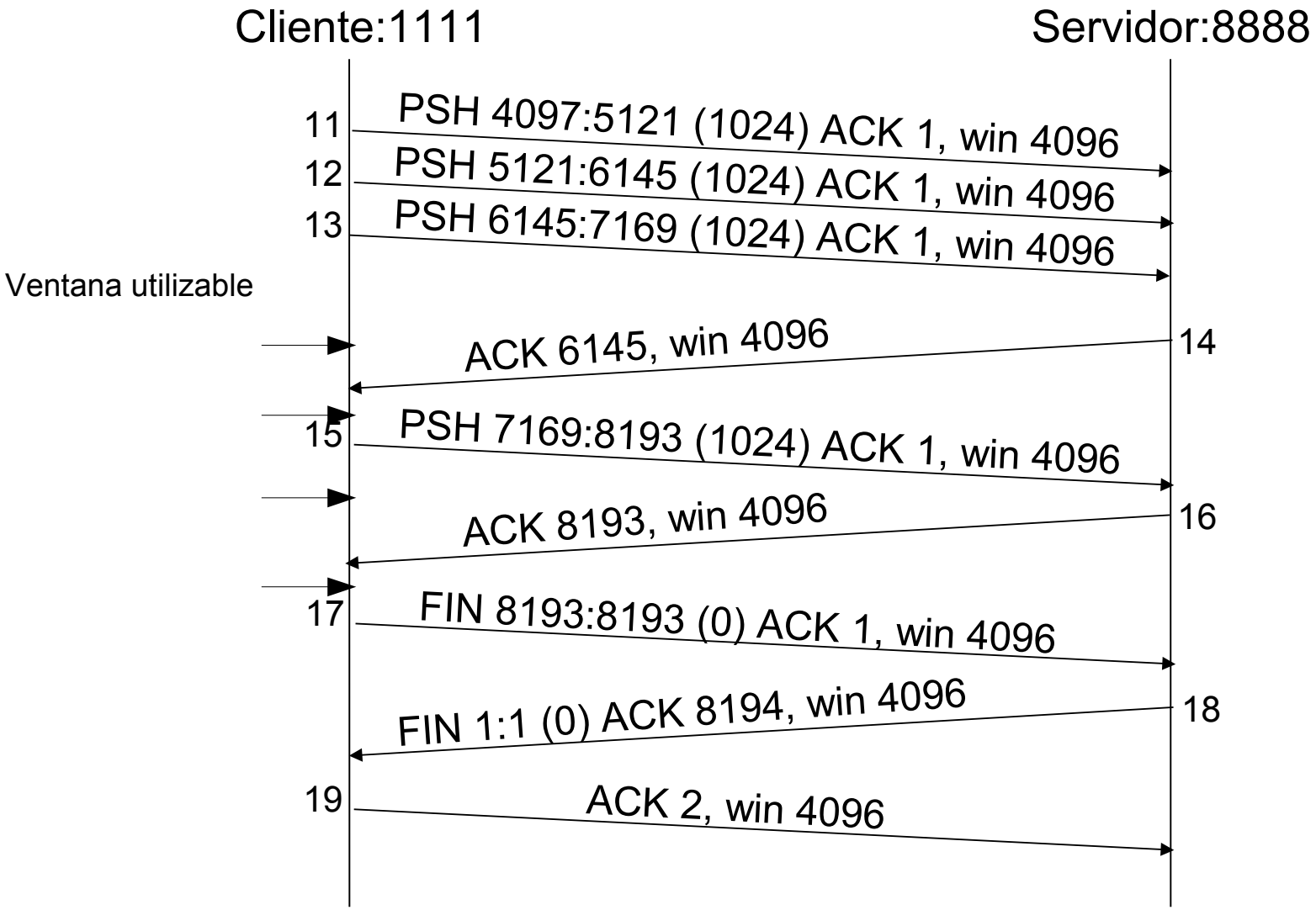
- Movimientos de la ventana:
  - Cierre (el eje izquierdo se desplaza a la derecha): se confirman datos (~ recibir un ACK nuevo).
  - Apertura (el eje derecho se desplaza a la derecha): el proceso receptor lee los datos confirmados y libera espacio en el buffer (~ win se mantiene).
  - Reducción (el eje derecho se desplaza a la izquierda): no se recomienda en TCP.



Control de flujo: Ejercicio



Control de flujo: Ejercicio



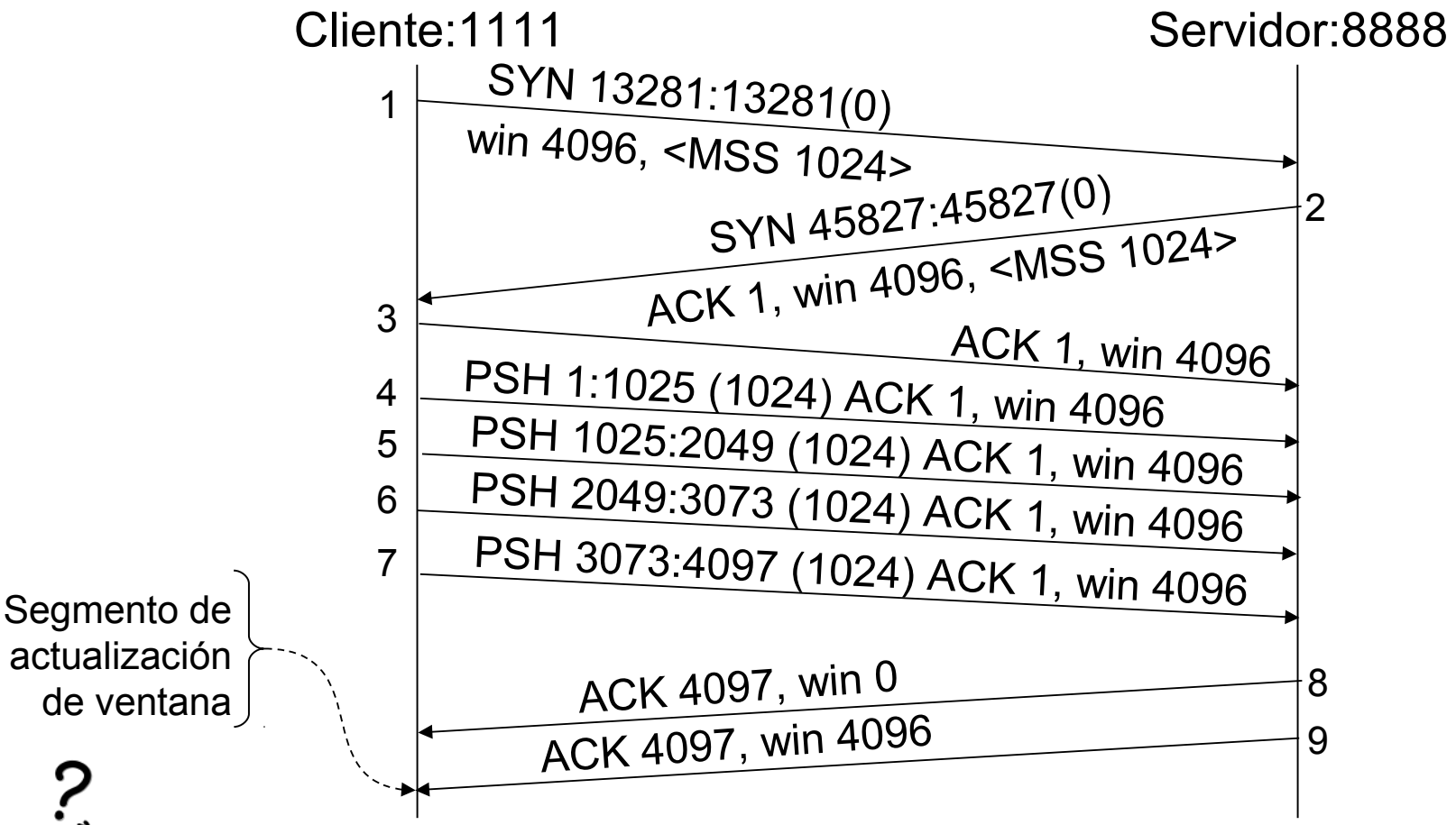
Control de flujo

- El envío de los mensajes TCP no es determinista, depende de múltiples factores: la carga de la red, la carga del receptor y emisor, …

- TCP utiliza ACKs acumulativos:
  - ACK 2049: confirma que se ha recibido correctamente hasta el 2048, y que el siguiente byte que se espera recibir es el 2049.
  - Segmentos 7, 14 y 16 son ACKs acumulativos.

- El emisor no tiene por que enviar una ventana completa de datos.
- El receptor no tiene que esperar a que se llene la ventana para enviar un ACK.

Control de flujo: Temp. de persistencia

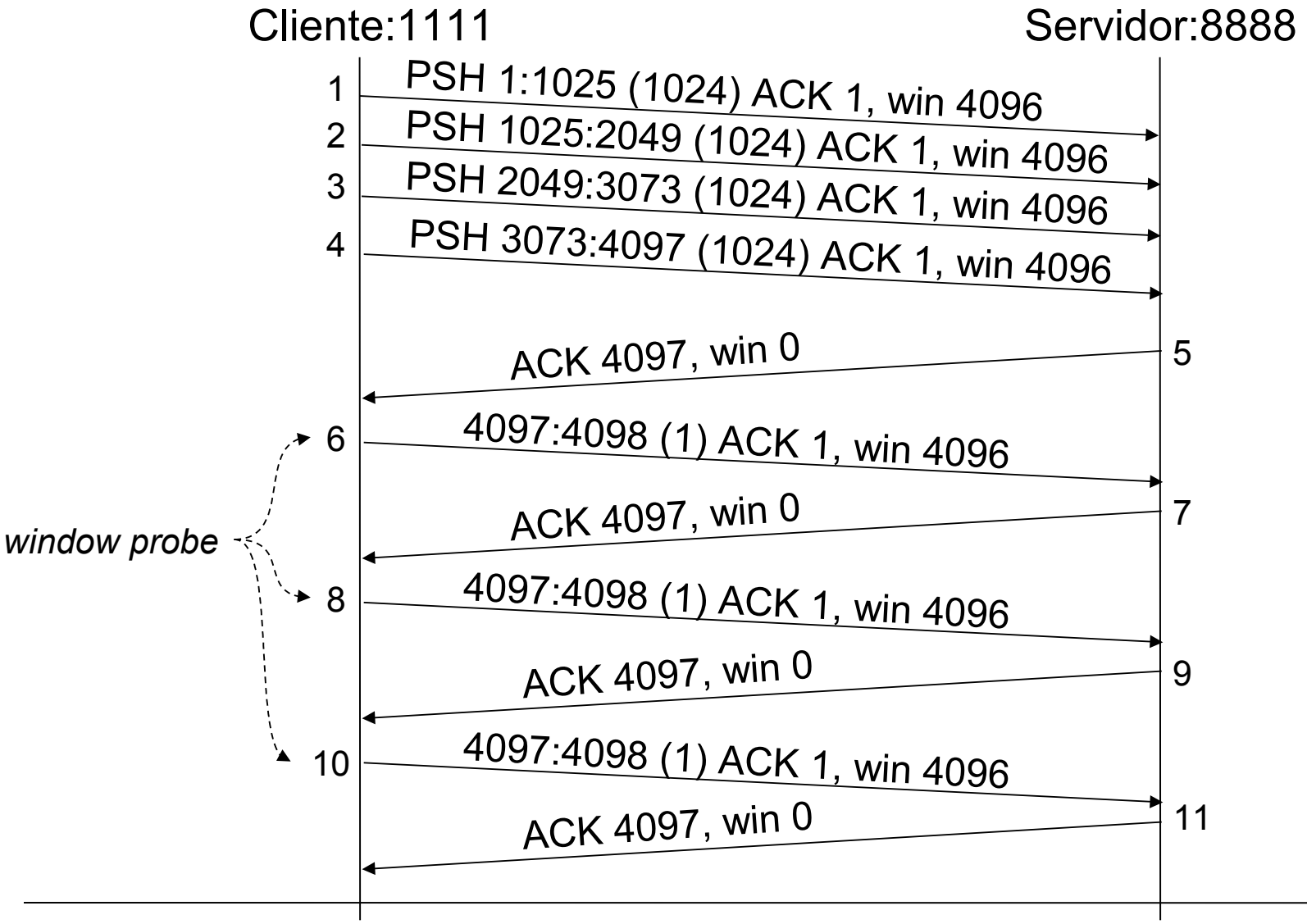


¿Qué pasa si se pierde el segmento de actualización de ventana?

Control de flujo: Temp. de persistencia

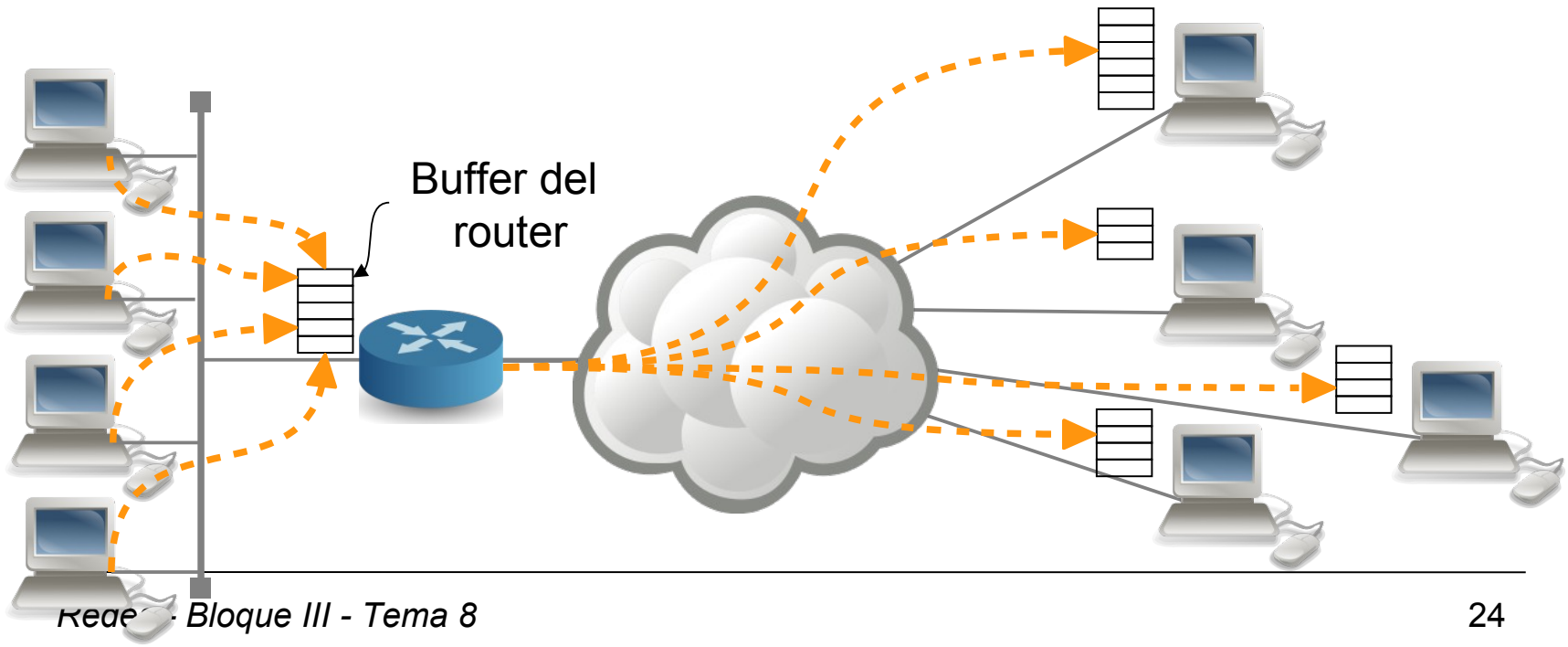
- → Se entra en una situación de abrazo mortal.
- Solución: **Temporizador de persistencia**.
  - Después de un tiempo sin que se abra la ventana, pregunta periódicamente si la ventana se ha actualizado utilizando unos segmentos especiales denominados: window probes.
- **Window probes:** segmentos de un byte utilizados para comprobar si realmente la ventana se ha modificado o no.
  - Se envía el byte de datos.
  - El receptor puede confirmar o no el window probe (en función del espacio libre en el buffer).
- El temporizador se activa siguiendo un *exponential backoff*. Se continúa retransmitiendo hasta que:
  - El receptor abre la ventana.
  - Se cierra la conexión TCP por las aplicaciones.

Control de flujo: Temp. de persistencia



Control de congestión

- El control de flujo evita que el emisor llegue a saturar al receptor.
- Esto es correcto en una LAN, sin embargo en un entorno con routers intermedios el control de flujo no es suficiente → Los routers intermedios también se puede saturar.
- Problema: los routers NO tienen nivel de transporte
- Solución: **Control de congestión**



Control de congestión

- Algoritmo de inicio lento: “La velocidad a la que se inyectan paquetes nuevos a la red es la velocidad a la que se reciben ACKs del otro extremo”.
  - Se inicia el envío de datos “poco a poco” → **Ventana de congestión**.
- Pero, ¿hasta cuándo? → Algoritmo para evitar la congestión:
  - Se establece el **umbral de inicio lento**.
  - Por debajo del umbral, se aplica el algoritmo de inicio lento.
  - Por encima, se aplica un crecimiento suavizado.

- http://media.pearsoncmg.com/aw/aw_kurose_network_4/applets/fairness/index.html

Control de congestión

- ¿Y qué pasa si algo va mal?

  1º: ¿Qué es que algo vaya mal? Cuando se pierde un paquete.
  - Dos posibilidades para “perder” un paquete:
    - Saturación en un router
    - Error en el paquete (probabilidad mucho menor del 1%)
  - Se asume que cuando se pierde un paquete es debido a que, al menos un router, está saturado.

  2º: ¿Cómo sé que algo va mal?
  - Se utilizan dos indicadores para identificar un problema de congestión (pérdida de un paquete):
    - Ha vencido un timeout de retransmisión.
    - Se han recibido ACKs duplicados.

  3º: ¿Qué hago cuando algo va mal?
  - Retransmitir.
  - Y reducir la velocidad de transmisión.

- Todo esto se integra en el **algoritmo para evitar la congestión**.

Control de congestión

- ¿Y si los routers pudiesen notificar que va a haber congestión? → Explicit Congestion Notification (**ECN**)
  - Deben ser routers avanzados (gestión de colas activa – AQM).
- Si un router está próximo a una situación de congestión → activa los bits de ECN en la cabecera IP.
- El receptor lo recibe y activa el flag ECN-Echo de la cabecera TCP en el ACK enviado.
- El emisor recibe el ACK con el flag ECN-Echo activado → Reacciona igual que si se hubiese perdido un paquete.
- Activa el flag CWR en los siguientes paquetes de datos → Notifica al receptor que ha recibido el aviso.

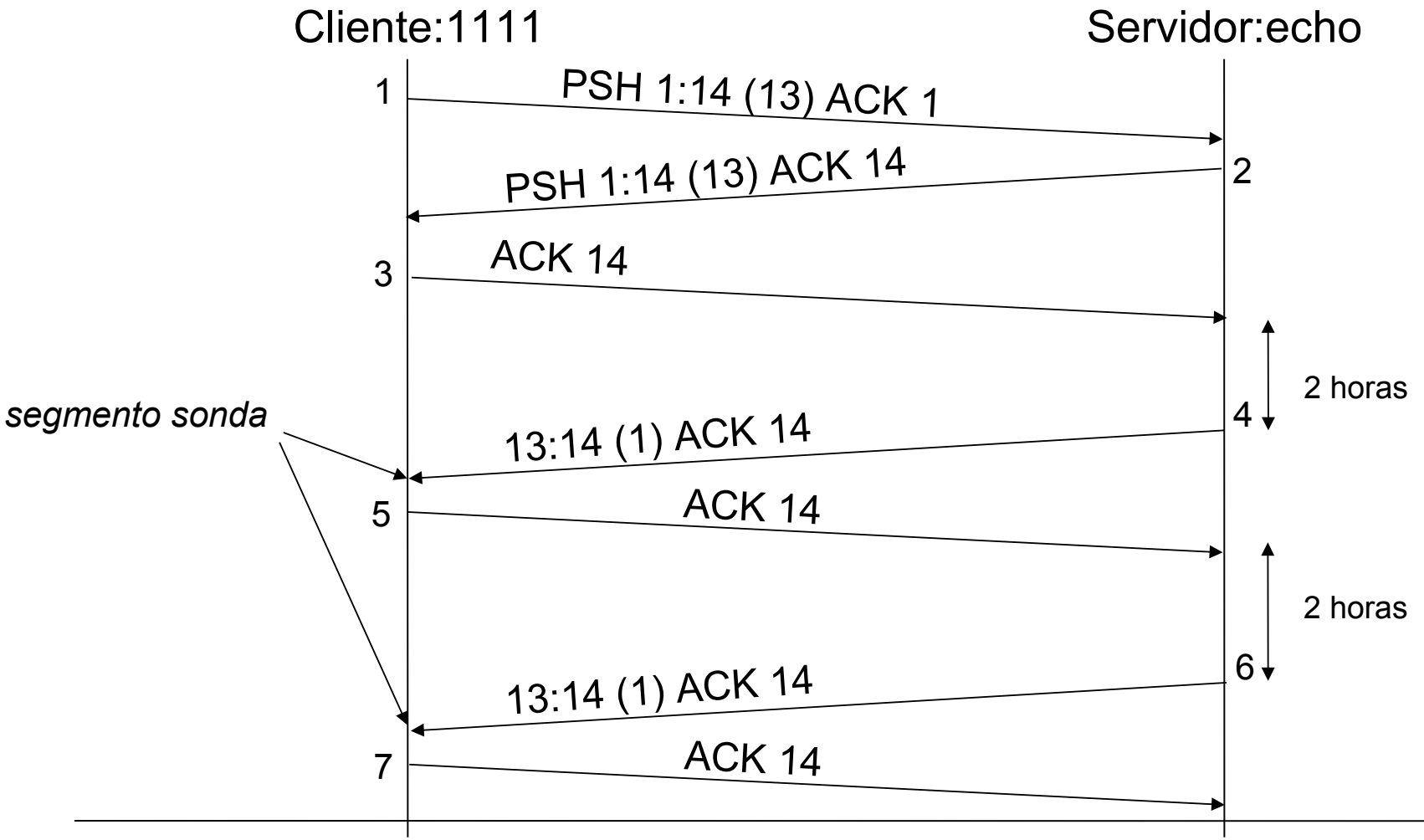
- Se controla con: /proc/sys/net/ipv4/tcp_ecn

Temporizador de keepalive

- En una conexión TCP sin intercambio de datos, no se produce ningún intercambio de paquetes (polling) → Problema en situaciones de fallos de uno de los extremos (normalmente el cliente).
- Solución: **temporizador de keepalive** → Mantiene la conexión activa (aunque no es parte del RFC de TCP).
- Permite liberar recursos al otro extremo, si realmente está desconectado.
- Funcionamiento: después de 2 horas de inactividad, el servidor enviará un segmento sonda (similar al window probe).
  - Segmento sonda: segmento de un byte, correspondiente al último byte enviado.
- El otro extremo puede estar en cuatro situaciones: ok, caído, reiniciado o no alcanzable.
- Se controla con (en /proc/sys/net/ipv4): tcp_keepalive_time (2h), tcp_keepalive_intvl (75 segs), tcp_keepalive_probes (9).

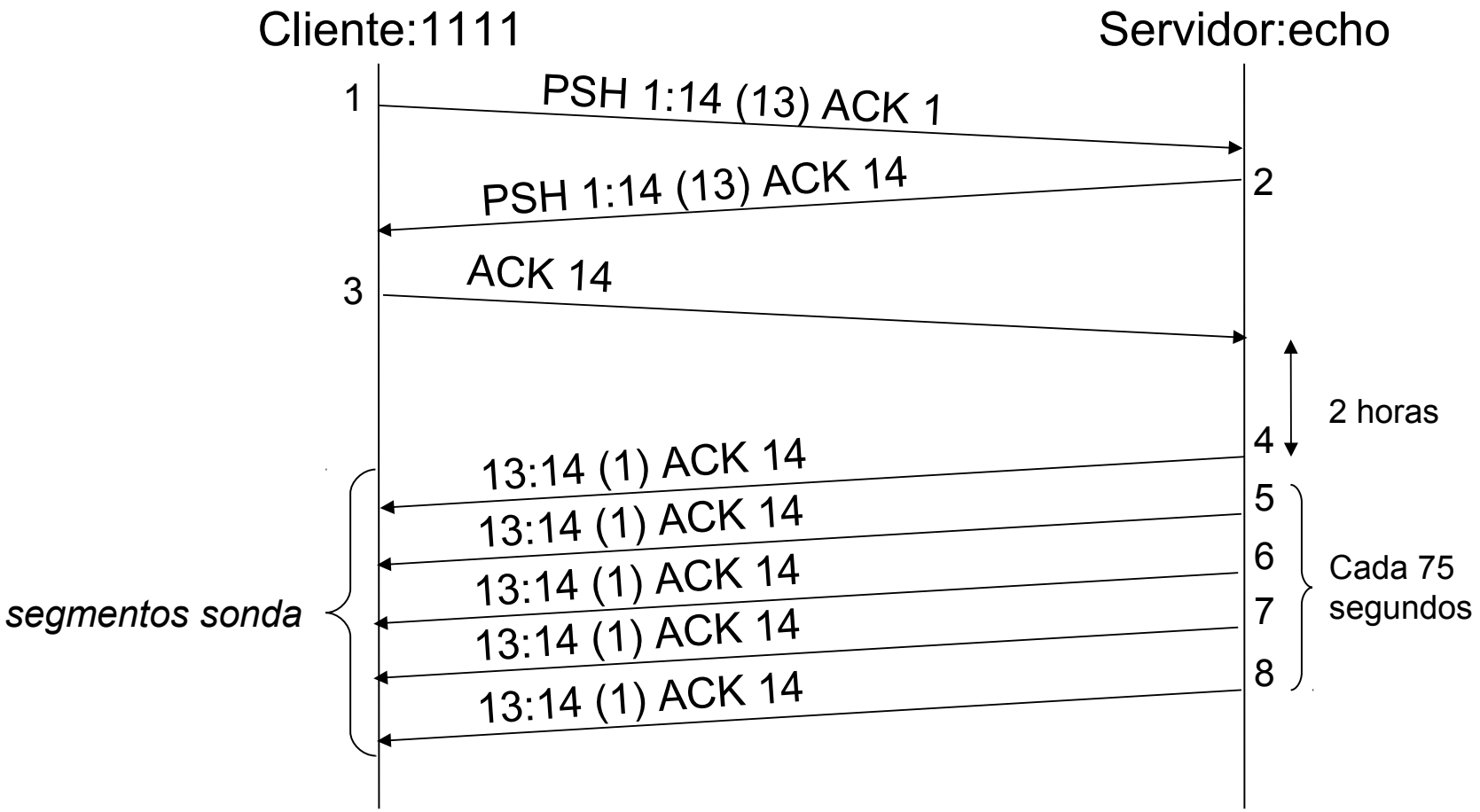
Temporizador de keepalive

- Caso 1: cliente OK



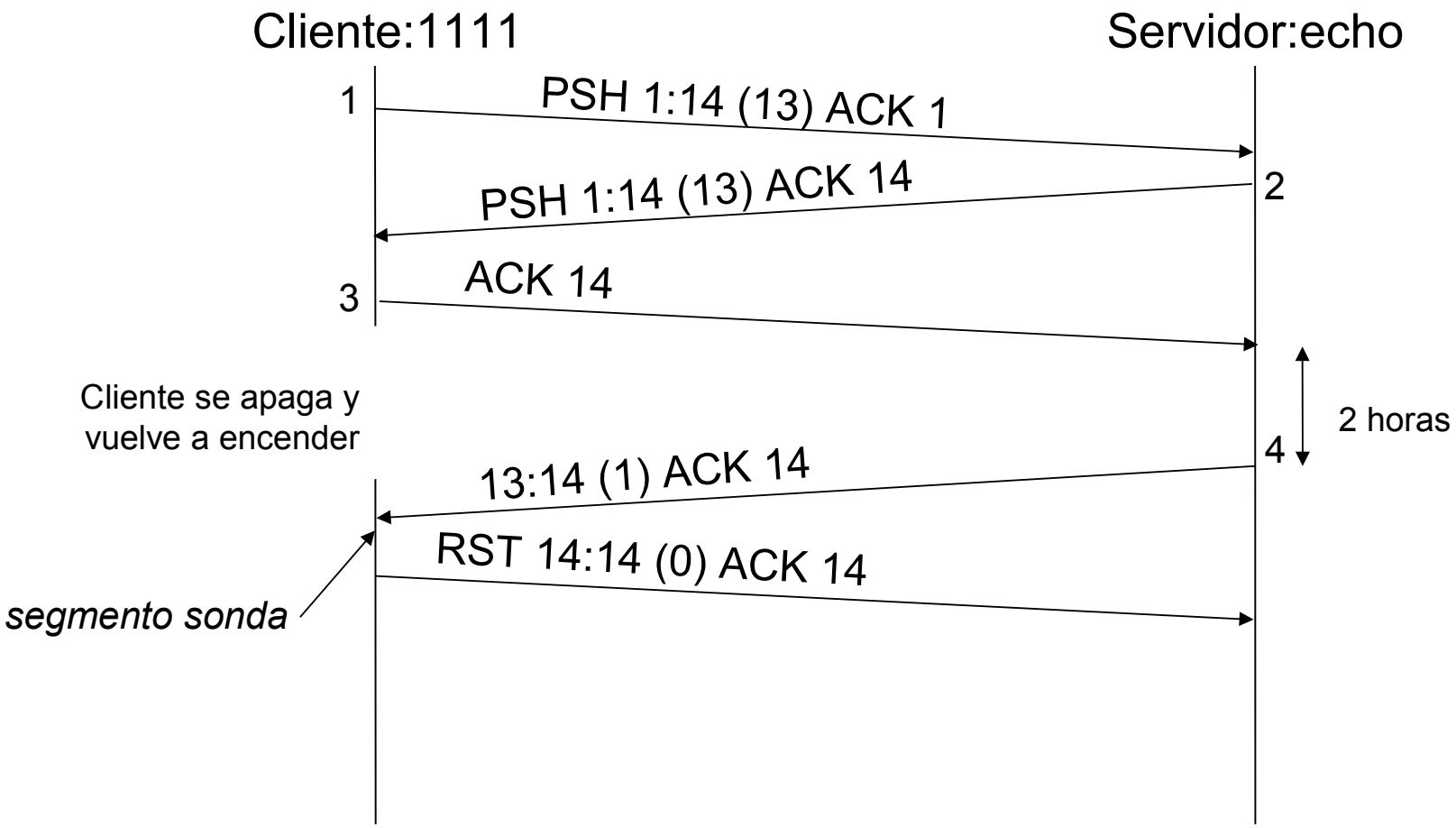
Temporizador de keepalive

- Caso 2: cliente caído



Temporizador de keepalive

- Caso 3: cliente caído y reiniciado



Temporizador de keepalive

- Caso 4: cliente inalcanzable (= caso 2)